

시나브로 배우는 자연어처리

바벨피쉬 송치성

신미리·민미나트·모래

국립·평생·직·명·행

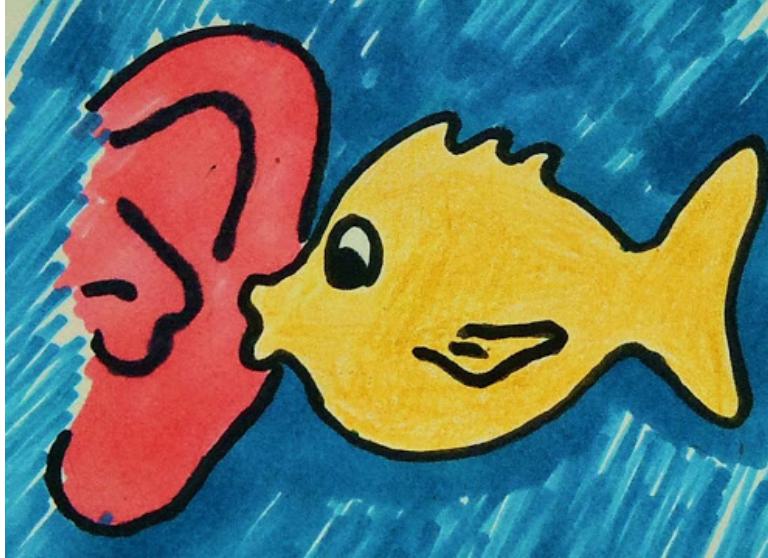
·안드·려·닐·샤·두·네

·보·논·가·너·그

스터디 소개

- 1. 스터디 소개**
- 2. 자연어처리**
- 3. 자연어처리 실습**
- 4. Word Embedding**

스터디 소개



바벨피쉬란?

- 더글러스 애덤스의 SF소설 '은하수를 여행하는 히치하이커를 위한 안내서'에 나오는 작은 물고기.
- 귀에 넣으면 어떤 언어로 이야기한것이든 즉시 이해할 수 있게 됨.
- 자료 및 커리큘럼 : <http://babelpish.github.io/>
- 페이스북 그룹 : <https://www.facebook.com/groups/babelPish/>

스터디 소개

모노가너거
인도·령나루·사·파
구·팡·장·직·민·향·
신·미·구·센·미·센·트·모·래

바벨피쉬py

복작복잡스핀

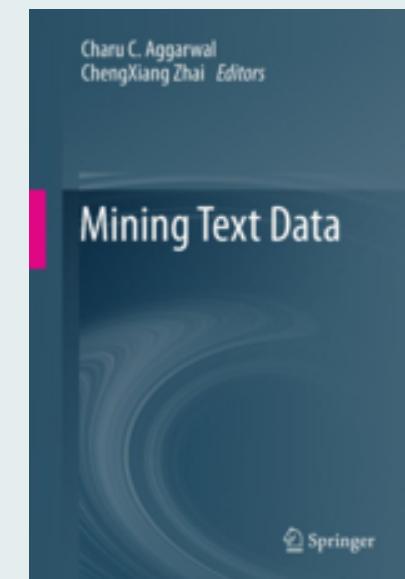
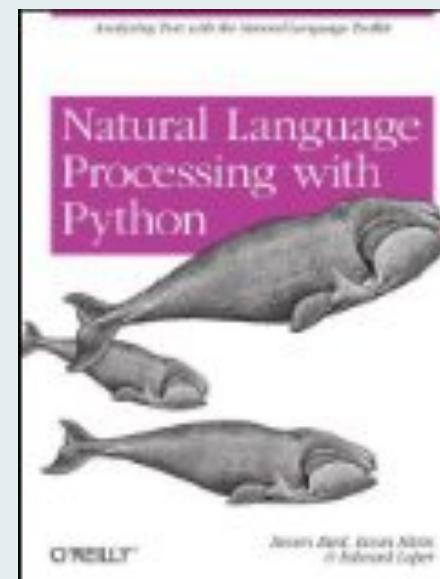
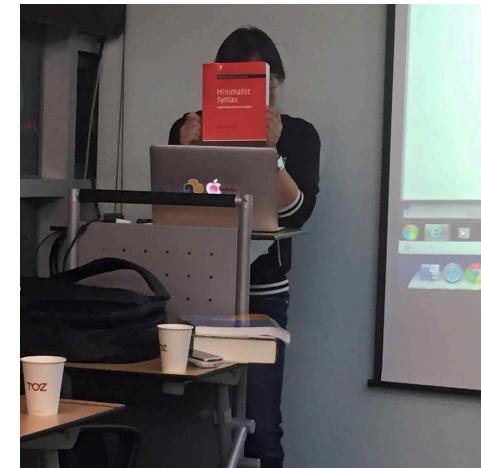


스터디 소개

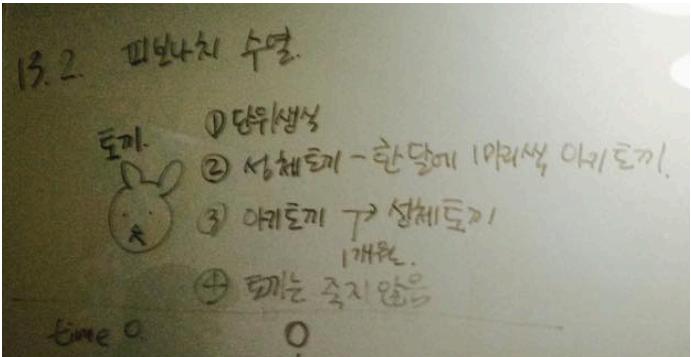
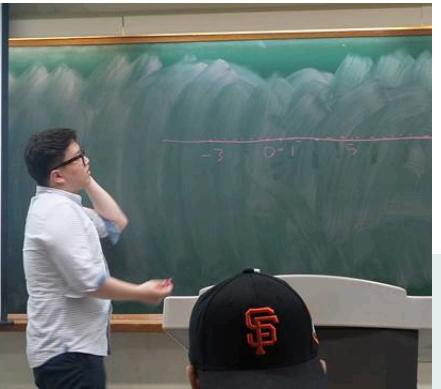
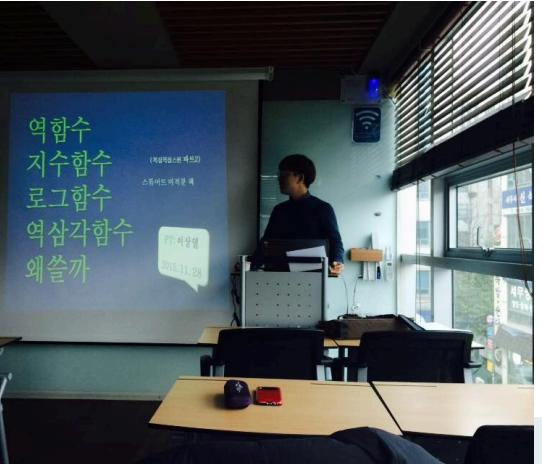
신미구는 미를 끌고 래
모는가 너 가
이드·려 날·사·되
귀·팡·장·직·명·행

바벨피쉬py

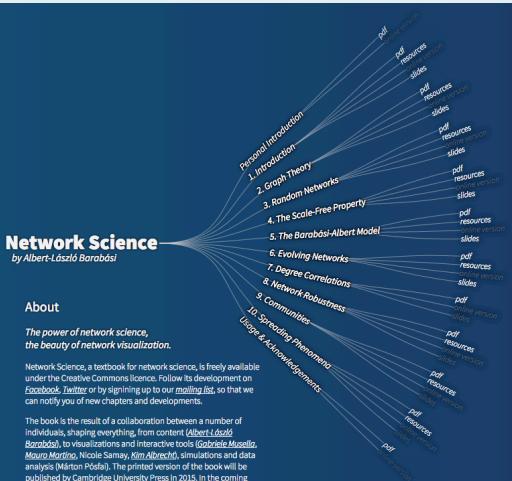
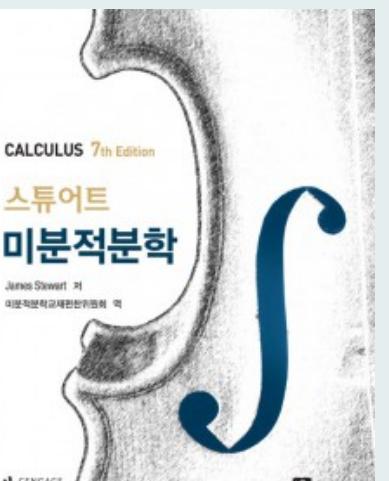
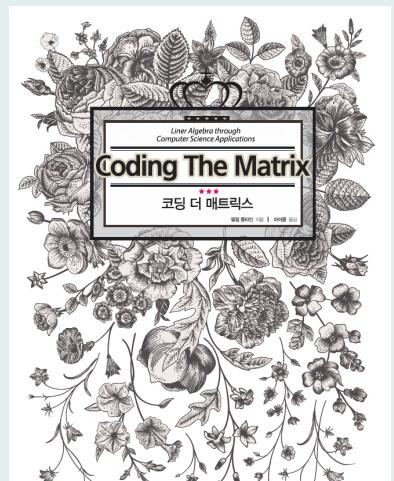
- 자연어처리 스터디.
- 비전공자도 서당개 체험.
- 재미있는 한글공부ㅋ



스터디 소개

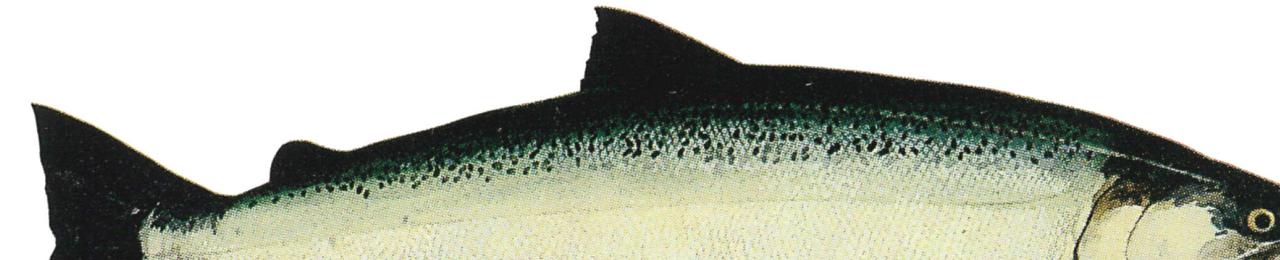
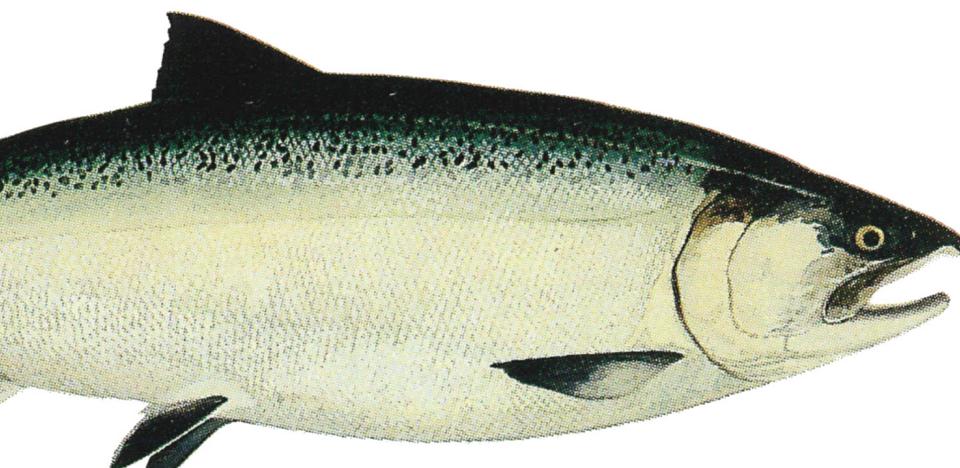
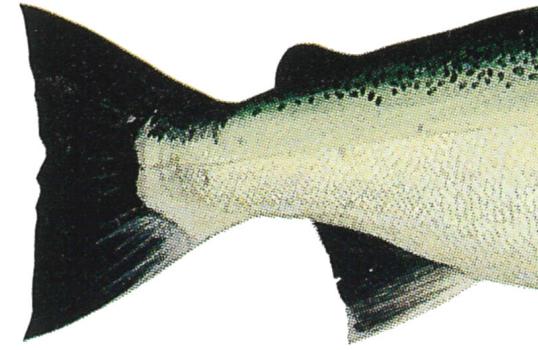
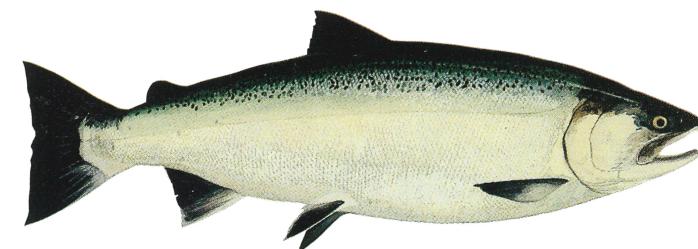
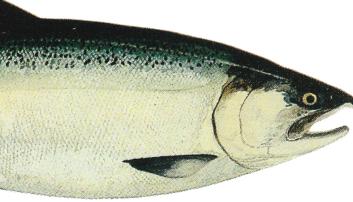


- 복잡계+수학 스터디
- 수포자도 할수있어요..!
- 사칙연산이 이렇게 어려웠나…



자연어처리

자, 연어처리...? 



자연어처리

언어 (Language) :

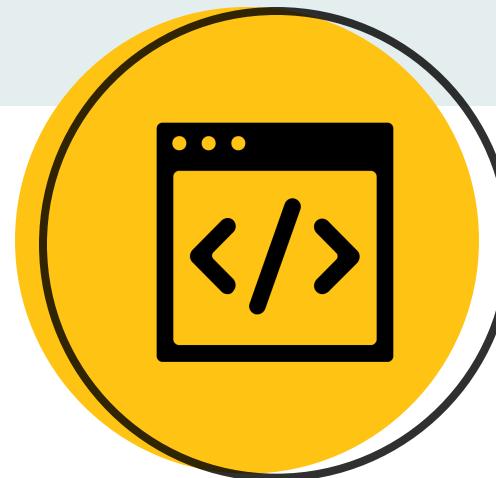
- 인간이나 동물들이 음성이나 문자 등을 사용하여 사상이나 감정을 나타내고 의사소통하는 수단.



사람의 언어



동물의 언어



프로그래밍 언어

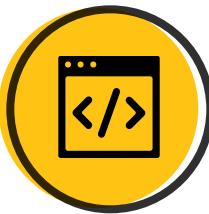
자연어처리

자연 언어 (Natural Language) :

- 의사소통을 위해 사용하는 언어와 같이 자연 발생적으로 생성된 언어. (↔ 인공언어)

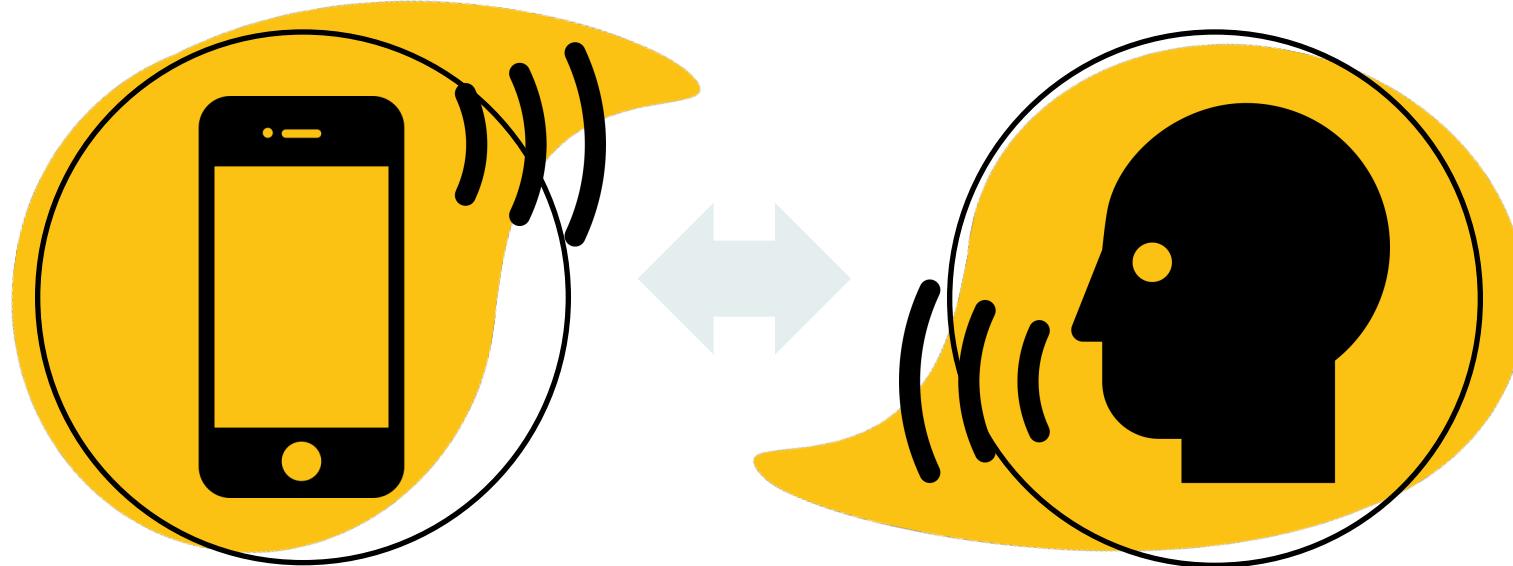


자연언어 : 한국어, 영어, 일본어



인공언어 : 프로그래밍 언어, 에스페란토어

자연어처리



자연어처리 (Natural Language Processing) :

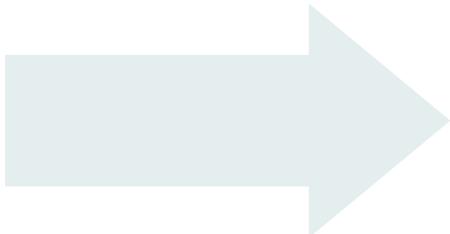
- 자연어를 분석하여 컴퓨터가 이해할 수 있는 형태로 만들거나 그러한 형태를 다시 인간이 이해할 수 있는 언어로 표현하는 제반 기술.

자연어처리

자연어 분석

- 형태소 분석
- 구문 분석
- 의미 분석
- 담화 분석
- 중의성 해소

뭘 할 수 있을까..?



응용 기술

- 검색
- 온라인 광고
- 자동번역
- 감정분석
- 음성인식
- 맞춤법검사

Cf) 구글이 하는것 : <http://research.google.com/pubs/NaturalLanguageProcessing.html>

자연어처리 실습

직접 해보자!

Step 1. NLTK 초간단 실습 : 영문으로된 텍스트를 형태소 분석해보기



Natural Language
Analyses with NLTK

자연어처리 실습

Step 1. NLTK 초간단 실습 : 영문으로된 텍스트를 형태소 분석해보기

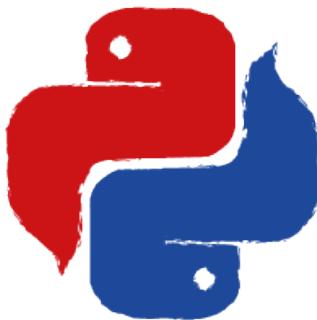
1. 문장 끝(EOS; End Of Sentence) 탐지
2. 토큰화(Tokenization)
3. 품사(POS; part-of-speech) 태깅(Tagging)

Jupyter notebook 링크 : <http://bit.ly/1R2WkIB>

자연어처리 실습

한글도 해보자!

Step 2. KonlPy 간단 실습 : 한글도 다뤄보자.



KoNLPy

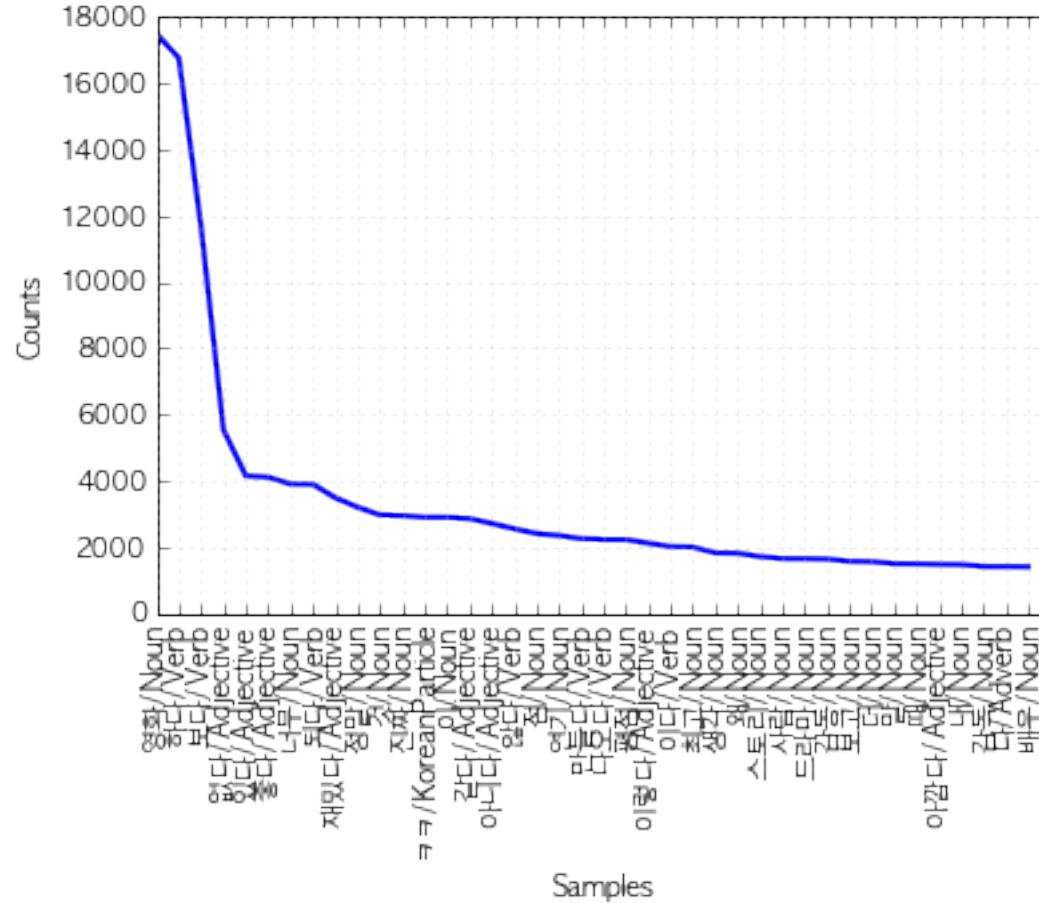
자연어처리 실습

Step 2. KonlPy 간단 실습 : 한글도 다뤄보자.

1. 한글 텍스트 데이터 불러오기
2. 트위터 형태소 분석기로 품사 태깅(POS Tagging)
3. 어떤 단어가 많이 사용되었는지 단어 빈도 플롯 살펴보기

Jupyter notebook 링크 : <http://bit.ly/1NSx0Rj>

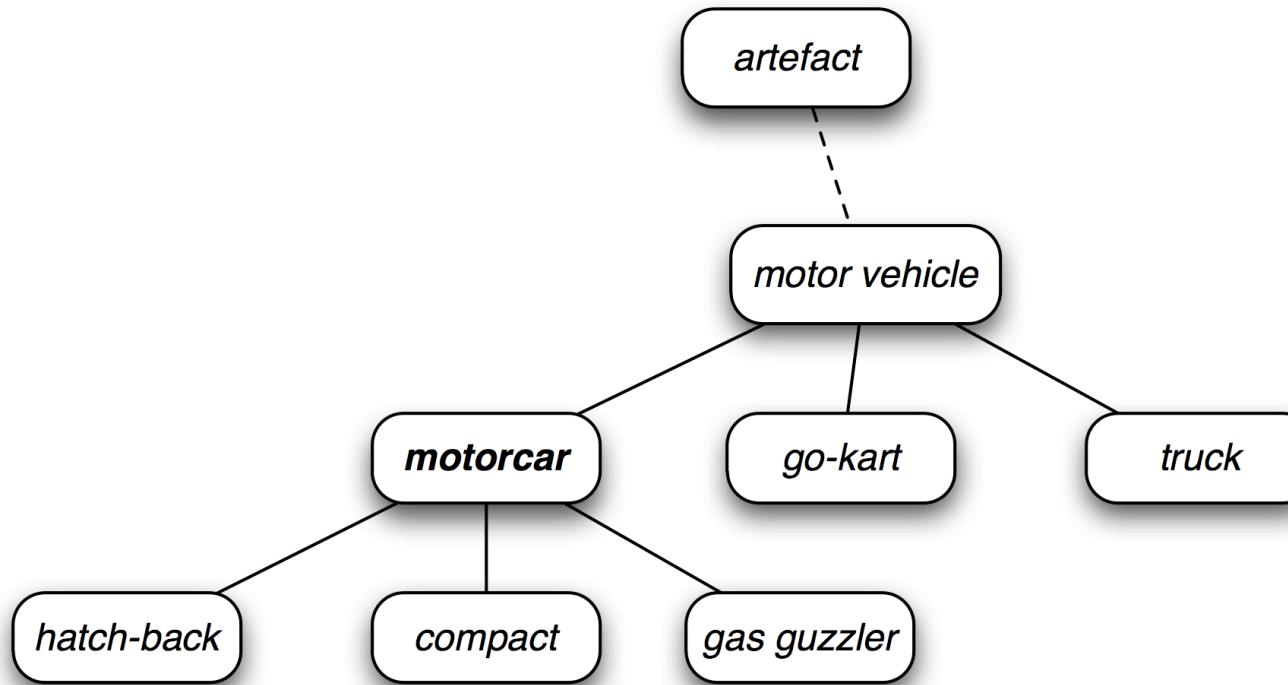
자연어처리 실습



자연어처리 실습

**이게 끌..?
분석은 어떻게 할까?**

자연어처리 실습



⟨Wordnet⟩

분류 체계(Taxonomy)를 분석하여 단어간 관계를 표현.

자연어처리 실습

좋아.
사랑해.
보고싶어.

싫어.
짜증나.
귀찮아.

분류 체계(Texonomy)를 이용하여 단어 의미(유사어) 파악.

자연어처리 실습

하지만.. 문제점.



신조어



뉴昂스



많은 노동력
필요

자연어처리 실습

- 대부분의 Rule-based / Statistical NLP에서는 형태소를 atomic symbol로 표현.
- 이때의 벡터표기는 이산적 표현(discrete representation)방식.

Motel [0 0 0 0 0 0 0 1 0 0 0 0 0]

Hotel [0 0 1 0 0 0 0 0 0 0 0 0 0]

Book [0 0 0 0 0 0 0 0 0 0 0 0 1]



- Sparse함.
- 단어들의 의미 유사성을 찾기 어려움!

<1-hot representation>

Word Embedding

Word Embedding을 이용하여 문제를 해결해보자!

Word Embedding

“You shall know a word by the company it keeps”

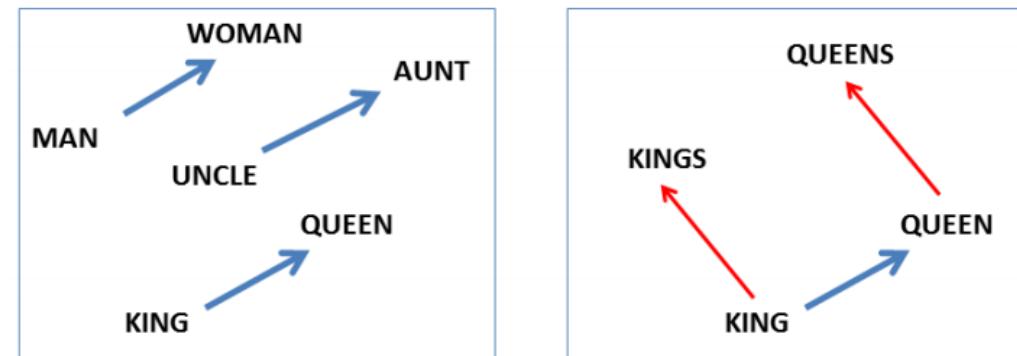
J. R. Firth 1957

별거 아닌거 같아도 이거 밤새서 만든거야.
이 식빵에는 밤이 많이 들어가서 맛있다.

Word Embedding

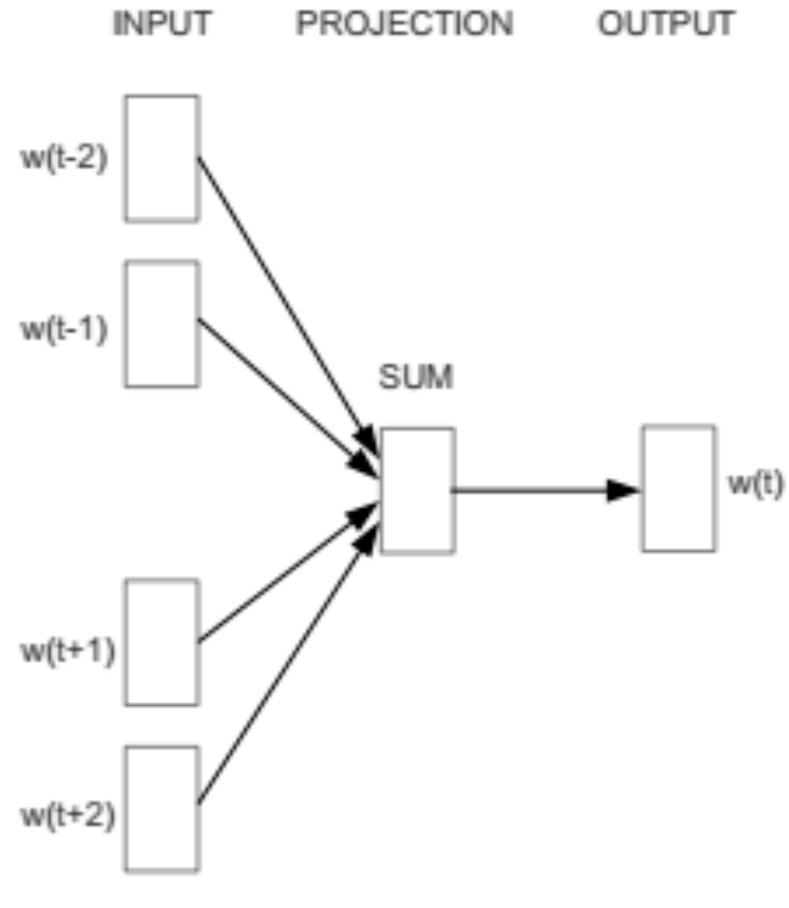
Word2vec

- Neural network를 이용하여 단어를 Distributed하게 표현.
 - Motel [0.23 0.34 ... 0.34 0.53]
 - Hotel [0.12 0.22 ... 0.12 0.23]
- Distributional Hypothesis : states that words that appear in the same contexts share semantic meaning
- 구글에서 2013년 발표.
- 효율적인 연산.
- 벡터연산이 가능.

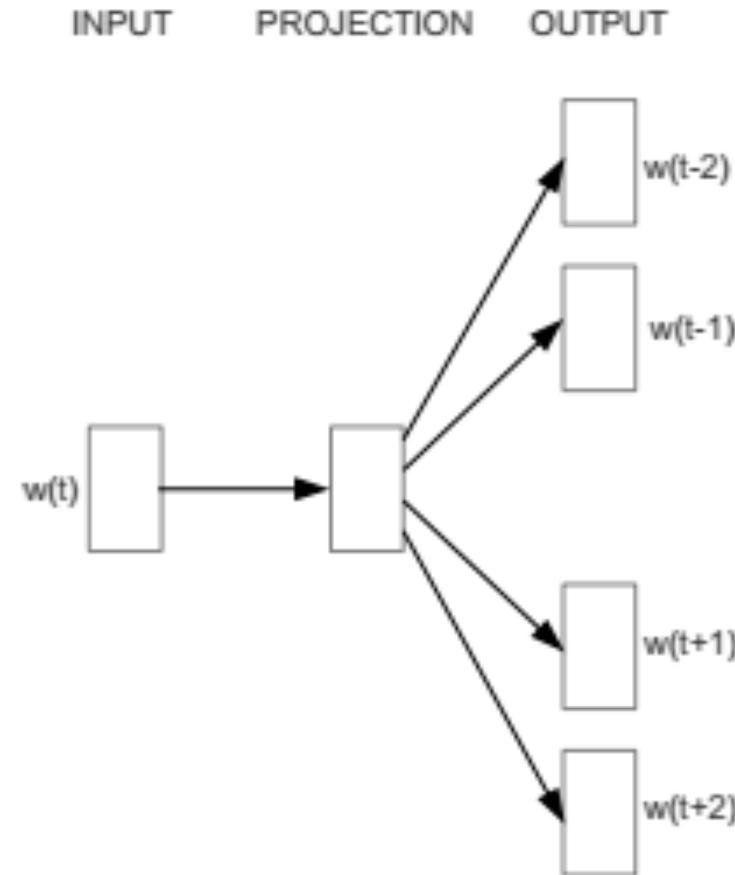


(Mikolov et al., NAACL HLT, 2013)

Word Embedding



CBOW

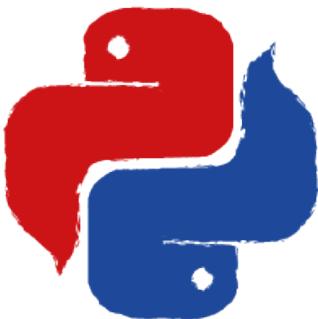


Skip-gram

Word Embedding



Step 3. Word2vec 실습 : 벡터 표현도 해보자.



KoNLPy

Jupyter notebook



자연어처리 실습

Step 3. Word2vec 실습 : 벡터 표현도 해보자.

1. Gensim을 이용하여 word2vec 모델링
2. word2vec 모델 갖고 놀기
3. t-SNE로 차원축소 및 Bokeh를 이용하여 시각화

Jupyter notebook 링크 : <http://bit.ly/1Ps5kHn>

자연어처리 실습

Word2vec 모델 테스트

```
# 만들어진 word2vec 모델 내에서 '취업'과 가까운(유사한) 상위 10개의 단어는?  
pprint(model_w2v.most_similar(positive=[u'취업/Noun'], topn=10))
```

```
[('취직/Noun', 0.8068336844444275),  
 ('졸업/Noun', 0.5646950006484985),  
 ('대학원/Noun', 0.5612005591392517),  
 ('진로/Noun', 0.5321317315101624),  
 ('대기업/Noun', 0.525084376335144),  
 ('진학/Noun', 0.5242314338684082),  
 ('스펙/Noun', 0.4993385076522827),  
 ('취준생/Noun', 0.48712578415870667),  
 ('취준/Noun', 0.4847239851951599),  
 ('공기업/Noun', 0.47719287872314453)]
```

```
# '끼부리다'와 가까운 상위 10개의 단어는?  
pprint(model_w2v.most_similar(positive=u'끼부리다/Adjective', topn=10))
```

```
[('철벽/Noun', 0.39050883054733276),  
 ('여우/Noun', 0.3863976001739502),  
 ('어장/Noun', 0.37865474820137024),  
 ('여자후배/Noun', 0.3712746798992157),  
 ('나쁜남자/Noun', 0.3340104818344116),  
 ('스킨십/Noun', 0.32896384596824646),  
 ('꼬리/Noun', 0.3282308578491211),  
 ('친한척/Noun', 0.3281872570514679),  
 ('나대다/Verb', 0.3261522948741913),  
 ('여후/Noun', 0.3258897066116333)]
```

자연어처리 실습

Word2vec 모델 테스트

```
# '미국'과 가까운 상위 5개의 단어는?  
pprint(model_w2v.most_similar(positive=u'미국/Noun', topn=5))
```

```
[(일본/Noun, 0.6754751801490784),  
(영국/Noun, 0.6259884834289551),  
(외국/Noun, 0.6108527183532715),  
(한국/Noun, 0.6050620675086975),  
(중국/Noun, 0.6048676371574402)]
```

```
# 다음 중 다른 하나는? (미국, 일본, 중국은 국가인 반면, 서울은 도시)  
pprint(model_w2v.doesnt_match(u'미국/Noun 일본/Noun 서울/Noun 중국/Noun'.split()))
```

서울/Noun

자연어처리 실습

Word2vec 모델 테스트

```
# 벡터연산 테스트.
```

```
# 어머니 - 엄마 = ??? - 아빠
```

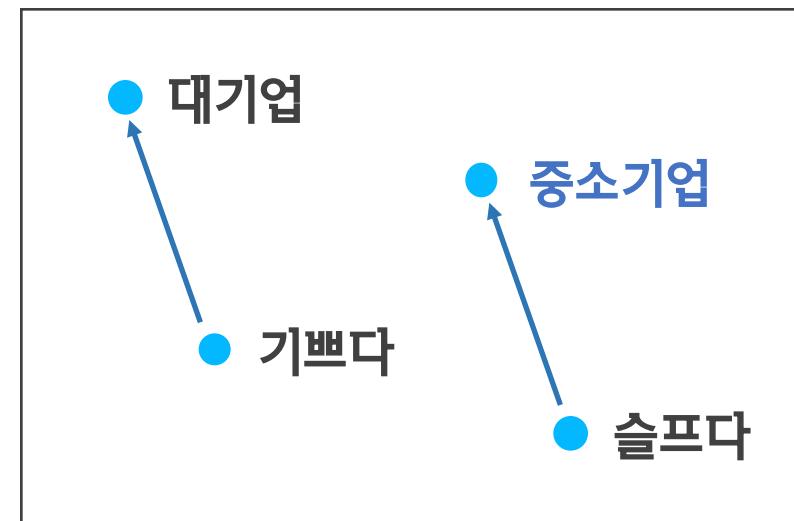
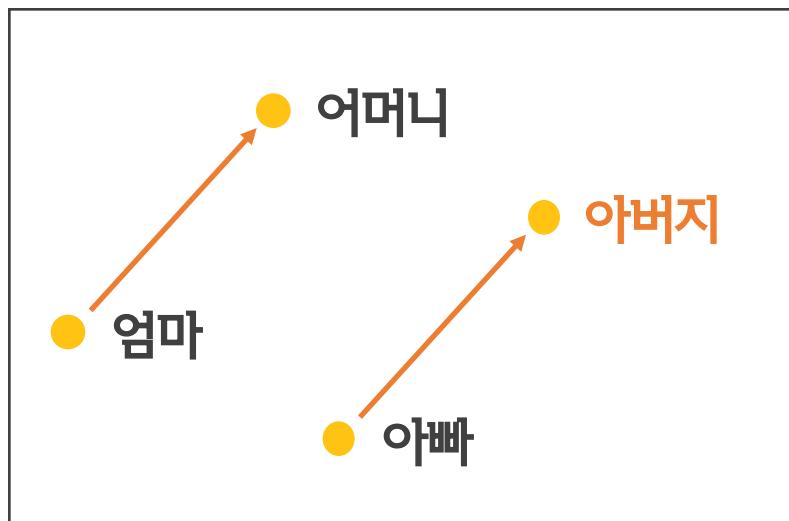
```
pprint(model_w2v.most_similar(positive=[u'어머니/Noun', u'아빠/Noun'], negative=[u'엄마/Noun'], topn=1))
```

```
[('아버지/Noun', 0.8476478457450867)]
```

```
# 대기업 - 기쁘다 = ??? - 슬프다
```

```
pprint(model_w2v.most_similar(positive=[u'대기업/Noun', u'슬프다/Adjective'], negative=[u'기쁘다/Adjective'], topn=1))
```

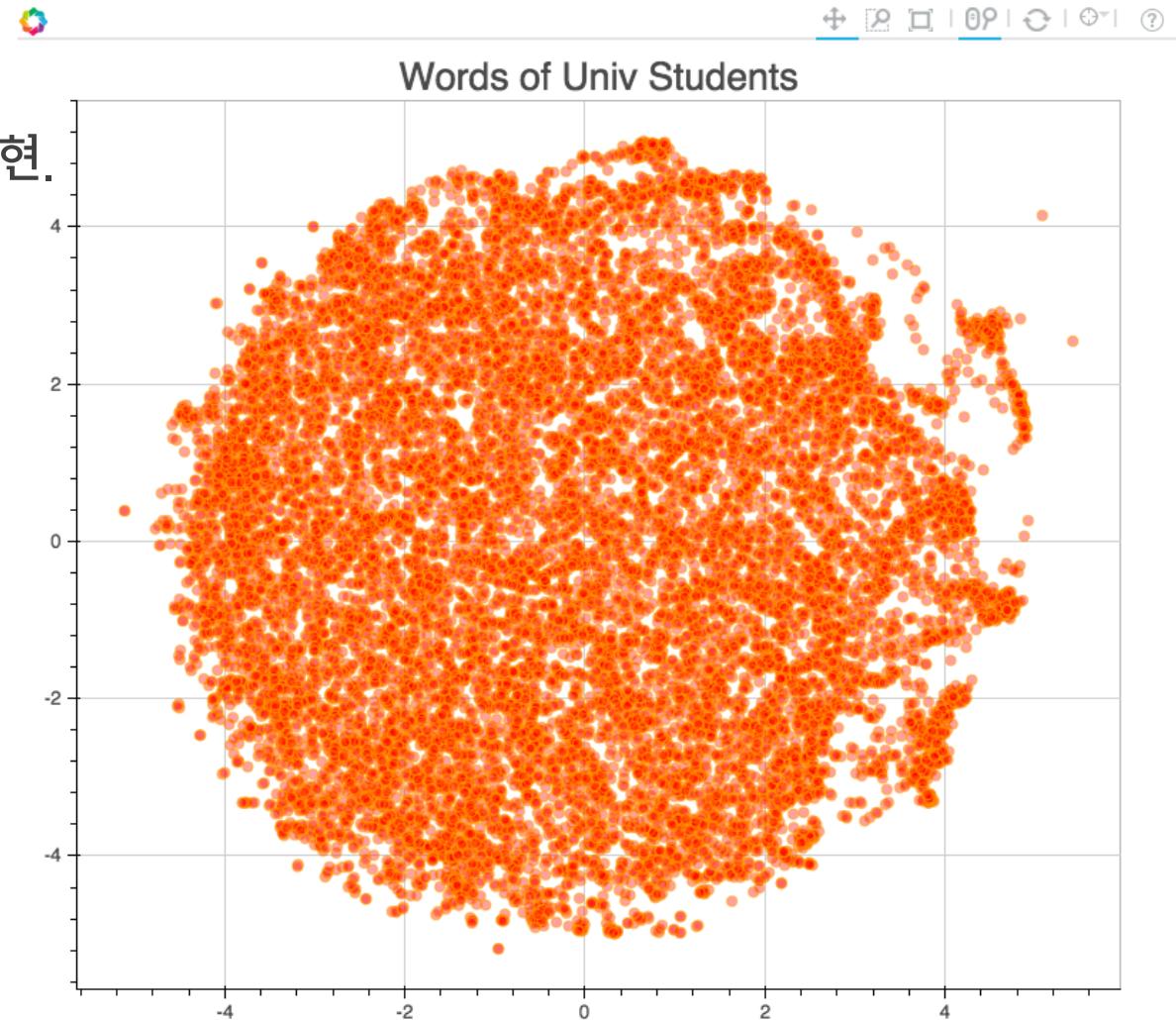
```
[('중소기업/Noun', 0.5176644325256348)]
```



자연어처리 실습

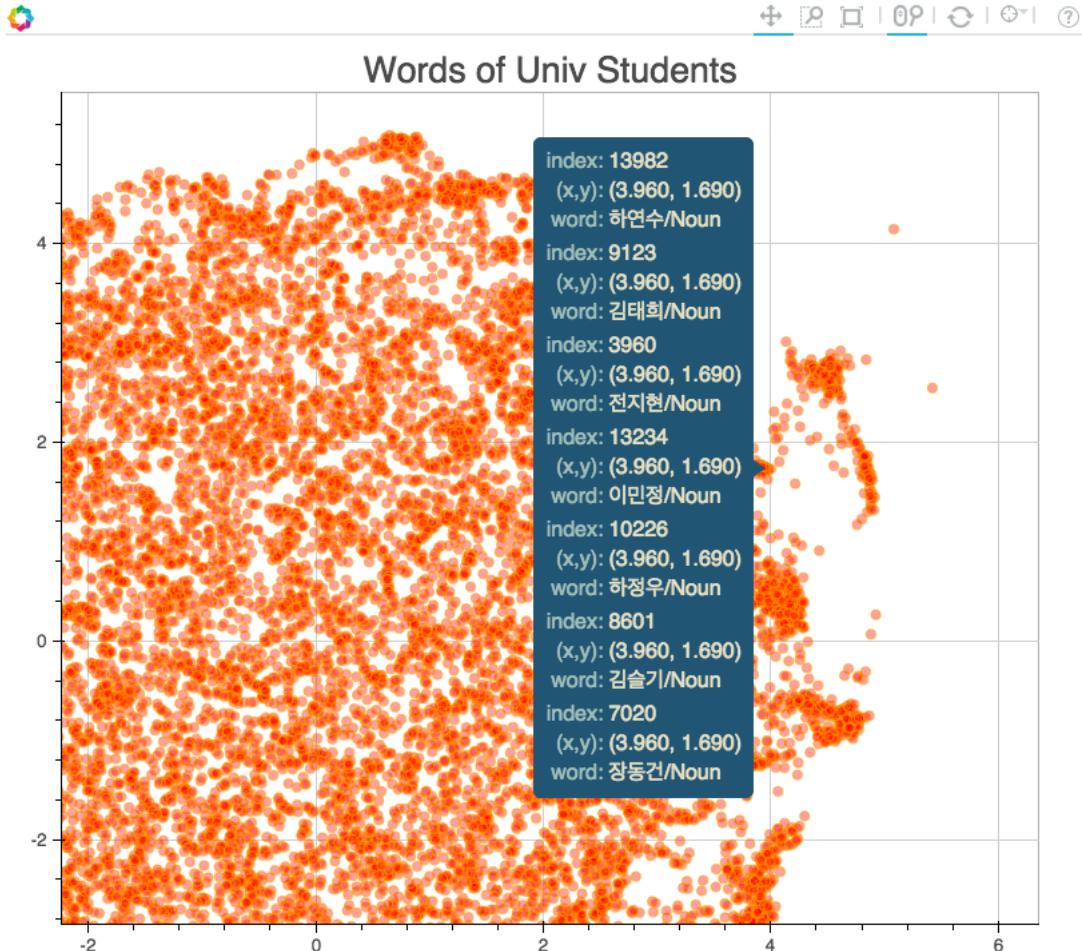
Bokeh를 이용하여 Word2vec 시각화

- 2차원으로 차원축소했기 때문에 2차원 좌표공간에 표현.
- 마우스인터랙션을 통해 데이터 탐색에 용이.
 - 마우스 오버 : 단어 정보 보기.
 - 마우스 드래그 : 좌표 공간 이동.
 - 휠 스크롤링 : 좌표 공간 주방.

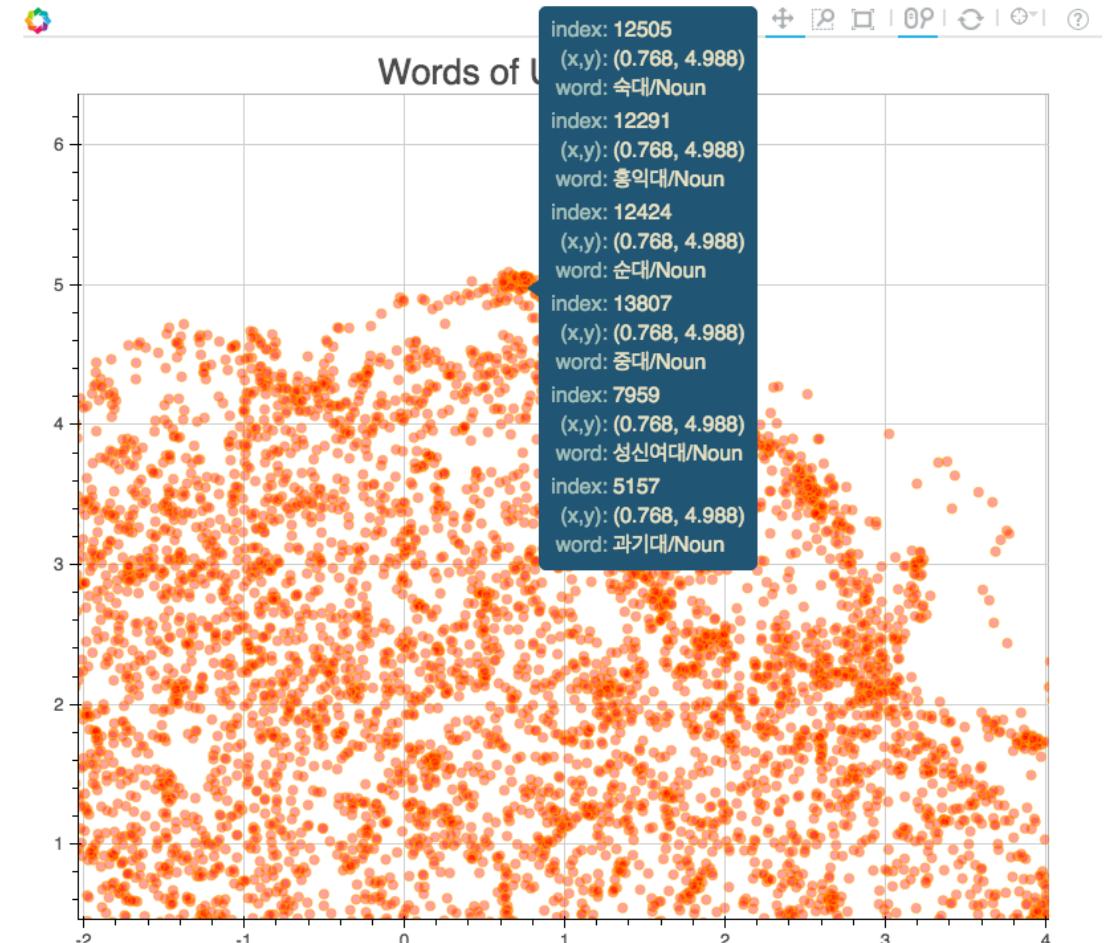


자연어처리 실습

Bokeh를 이용하여 Word2vec 시각화



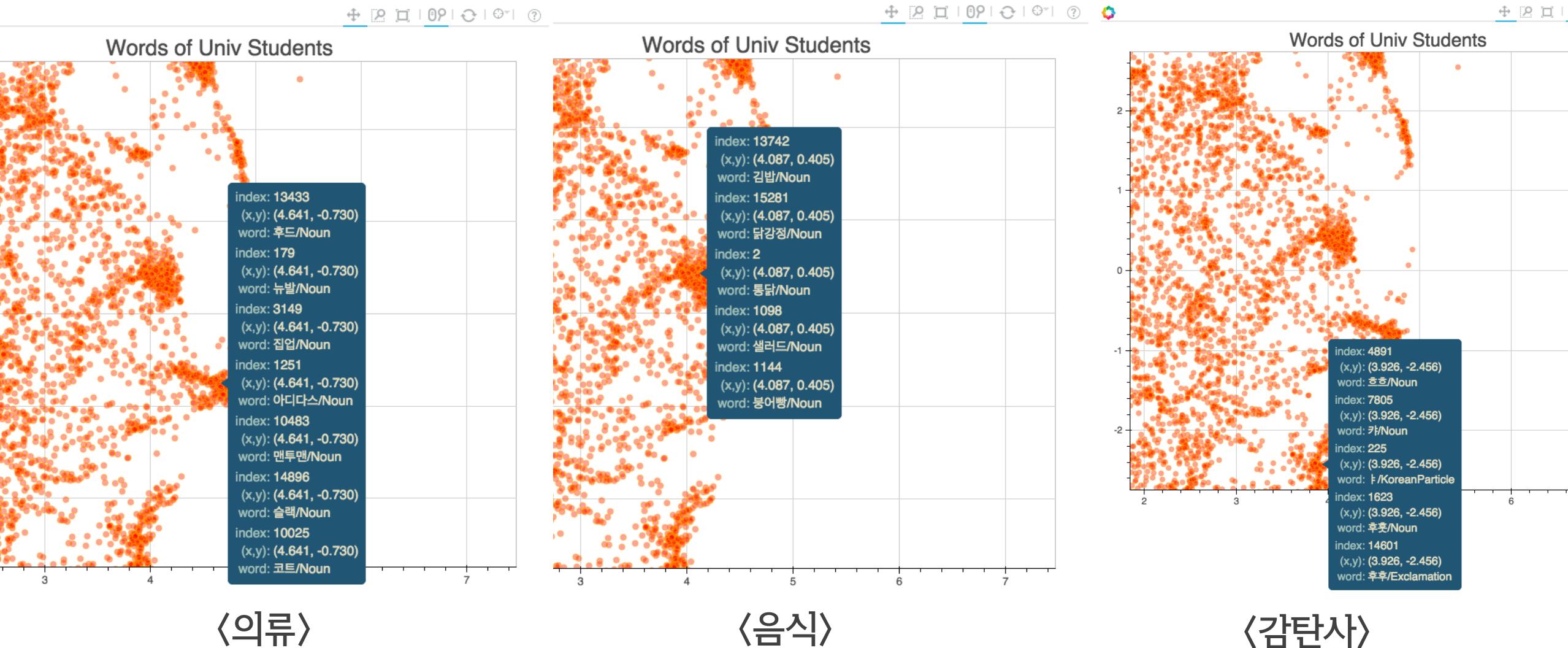
〈연예인〉



〈대학교〉

자연어처리 실습

Bokeh를 이용하여 Word2vec 시각화





E-Mail : daydrilling@gmail.com

Facebook : www.fb.com/shuraba